Machine Learning Engineer Nanodegree
Capstone Proposal

Abdelrahman Rashwan
April 9th, 2020

Bertlesmann Arvato Customer segmentation project

Domain Background

Bertelsmann found its origins as a publishing house in 1835 (Schuler, 2010), and through steady growth and development made its way to the software and hardware distribution market in the 80's (Computerwoche, 1983). By 1999 the company received its current name Arvato Bertelsmann (Neuer Name, neue Ziele, 1999) and over the next decade fully entered the domain of high-tech, information technology, and e-commerce services (Paperlein, 2012).

Arvato offers financial solutions in different forms , from payment processing to risk management. It is in this domain that that this capstone project will be developed. Arvato is looking to use its available datasets to support a mail order in identifying the best data driven way to acquire new client base. To achieve this goal Arvato's existing datasets will be explored to identify attributes and demographic features that can help segment customers of interest for this particular client.

Problem Statement

The problem statement for this project is "How can the mail order company acquire new clients in an efficient manner?".

The solution proposed for this problem is divided to 3 sections.

An unsupervised learning approach will be used to identify customer segments of value based on demographics data of existing customers versus general population data, after the segmentation has been done , a supervised learning technique will be used on the dataset containing data of customers that were a target of the marketing campaign in order for a model that predicts whether or not a person will turn into a customer to be built and finally using said model to turn in predictions on a final dataset for potential customers in order to determine which ones are worth being added to the marketing campaign list.

Datasets and Inputs

All the datasets were provided by Arvato in the context of the Udacity Machine Learning Engineer Nanodegree, on the subject of Customer Acquisition / Targeted Advertising prediction models.

There are 4 datasets to be explored in this project:

- Udacity_AZDIAS_052018.csv: Demographics data for the general population of Germany; 891 211 persons (rows) x 366 features (columns)
- Udacity_CUSTOMERS_052018.csv: Demographics data for customers of a mail-order company; 191 652 persons (rows) x 369 features (columns)
- Udacity_MAILOUT_052018_TRAIN.csv: Demographics data for individuals who were targets of a marketing campaign; 42 982 persons (rows) x 367 (columns).
- Udacity_MAILOUT_052018_TEST.csv: Demographics data for individuals who were targets of a marketing campaign; 42 833 persons (rows) x 366 (columns).

And 2 metadata files associated with these datasets:

- DIAS Information Levels — Attributes 2017.xlsx: a top-level list of attributes and descriptions, organized by informational category
- DIAS Attributes — Values 2017.xlsx: a detailed mapping of data values for each feature in alphabetical order

Solution Statement

For a two stepped problem a two stepped solution is proposed.

Since the first portion of the solution requires the usage of unsupervised learning methods selection and encoding any non-numerical features, followed by removal of highly missing columns, followed by imputation of any residual missing data will be done , followed by feature scaling to guarantee that the natural scale of the features does not affect their overall weight on the principal components, PCA for dimensionality reduction and as a part of the data pre-processing will be used and for the prediction step both GaussianMixture and KMeans

will be implemented as forms of clustering and performances will be compared.

Once the data is pre-processed and the customer segments are identified I will approach the supervised learning component of this project by testing which models, out of the considered habitual options for customer conversion prediction work best for these particular datasets, namely:

- XGBOOST
- RandomForestRegressor and GradientBoostingCLassifier)
- LGBM
- GridSearchCV
- RandomizedSearchCV

Benchmark Model

For this problem it is suggested to use Gradient Boosting Classifiers as they are considered state of the art and also based on data sets of relevance on Kaggle relating to customer conversion and churn prediction.

Evaluation Metrics

For the first part of the problem using unsupervised learning, explained variance ratio can be used when we are implementing PCA, as it accounts for the description of feature variance.

For the clustering part of the project (unsupervised learning) I will use Sum of squared error(SSE) as a metric on how many clusters should be implemented however this is not a clear indication for good clustering , an additional silhouette score will b calculated as it is a good representation for how good the clustering process is , however a full silhouette score analysis will be challenging as the we are dealing with such a large dataset therefore computation time will be ''unreasonable'' , Therefore, silhouette score will only be used to compare between the Kmeans and GMM models and then the model with the least SSE will be chosen

For the prediction part of the project (supervised learning), as data is highly unbalanced , none of the conventional metrics as accuracy can be considered and in the cases of unbalanced data either 'AUC-ROC' score or 'MCC' score are reliable.

'AUC-ROC' score will be taken as a metric as it is widely used and implemented in ready to go libraries

Project Design

1. **Data Cleanup:** most of the data that is received raw requires an extensive step of cleanup for improper data entries and missing values. For each feature I will examine the percentage of missing values, identify outliers and the type of feature (categorical, numerical). Missing data will dropped if a certain threshold is met or filled on a case by case approach.

2. **Data Visualization:** Allows for a birds-eye view of the data and early detection of particular patterns, namely, correlations between predictors and target variables, ranges and scales. For this we can take advantage of the matplotlib library and seaborn as well as pandas for preliminary summary statistics.

3. **Feature Engineering:** Implement PCA, find most relevant features, eliminate features of low importance for optimal model training further in the project.

4. **Model Selection:** Experiment with the before-mentioned algorithms to find the ideally suited for this problem, namely KMeans and GMM for the unsupervised learning portion and LGBM, RandomForestRegressor ,GradientBoostingCLassifier , XGBOOST, GridSearchCV and RandomizedSearchCV for the supervised learning portion in which we are to predict costumer conversion through targeted campaigns.

5. **Model Tuning:** Once we find the model that best suits our data, adjust model parameters within a range that allows for increased performance without overfitting.

6. **Test and Predict:** use the previously proposed metrics, explained in the table present in the section for evaluation metrics as an indicator of success in our predictions.

Reference

- Computerwoche. (1983, October 21). Bertelsmann vertreibt Rechner von TI. *Computerwoche*. Quora